Data Engineering Project -2

Azure Data Monitoring For Tokyo Olympics Data

Team:

Aathirainathan P
Ambati Sesha Sai Sahithya
Eswara Venkata Sai Raja

Date: 19-12-2024

---

1. Project Overview:

The Data Quality and Monitoring project leverages Azure's data engineering tools to build a system that ensures high-quality, trustworthy datasets. By integrating Azure Data Factory (ADF) and Azure Databricks, the project monitors, validates, and reports on 10 data quality metrics for the Tokyo Olympics dataset. This ensures robust workflows, enabling seamless analytics and data-driven decision-making.

2. Project Statement:

This project focuses on building an automated and scalable data pipeline to:
1. Ingest data from multiple sources (e.g., GitHub) into Azure Data Lake Storage.
2. Perform data quality checks and validations to identify and rectify inconsistencies.
3. Store quality metrics for reporting and decision-making.
4. Enable proactive monitoring with alerting mechanisms for data quality issues.

3. Data Overview:

The dataset for this project contains data from the Tokyo Olympics, including details about athletes, teams, coaches, gender participation, and medal counts. Key files include:

- *Athletes.csv*: Athlete details like Name, NOC (National Olympic Committee), and Discipline.
- *Coaches.csv*: Coaches' details and associated events.
- *EntriesGender.csv*: Gender-based participation metrics for various events.
- *Medals.csv*: Medal counts and rankings by country.
- *Teams.csv*: Team event participation information.

Schema Details:

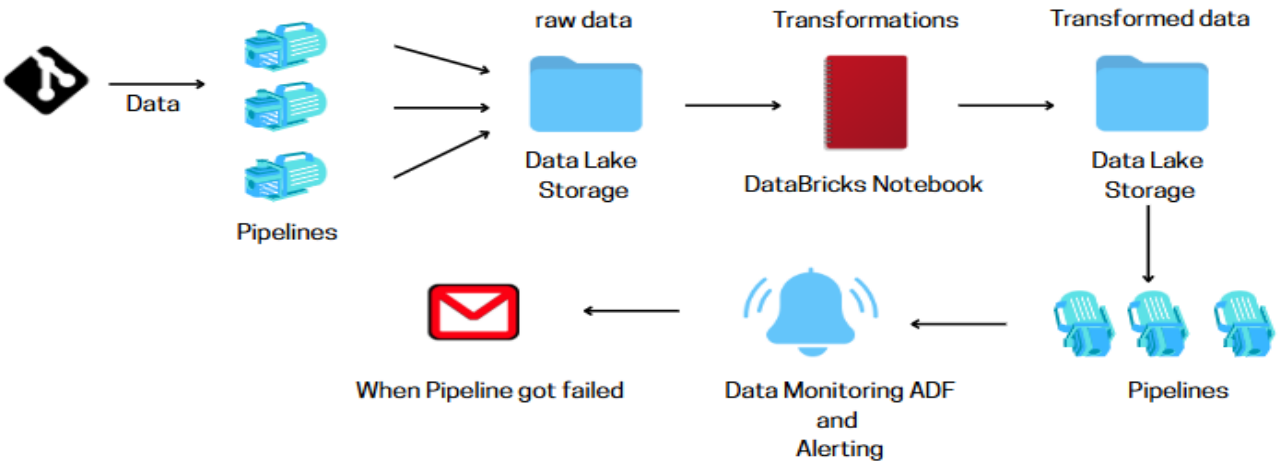
Column Name	Description
Athlete ID	Unique identifier for each athlete.
Name	Name of the athlete.
NOC	National Olympic Committee associated with the athlete.
Discipline	Sports discipline in which the athlete participates.
Coach ID	Unique identifier for each coach.
Coach Name	Name of the coach.
Event	Event associated with the coach or athlete.
Gender	Gender of participants (e.g., Male, Female).
Team	Team associated with the athlete or coach.
Medal	Medal won by the athlete (e.g., Gold, Silver, Bronze).
Rank	Rank of the team or athlete in the event.
Total Participants	Total number of participants in the event.
Category	Category of the event (e.g., Individual, Team).
Region	Geographical region of the participant's NOC (e.g., Asia, Europe).
Country	Country of the athlete or coach.
Event ID	Unique identifier for each event.
Participation Count	Number of participants in each gender category.
Null Count	Count of null or missing values in specific columns.
Duplicate Count	Count of duplicate records identified during quality checks.
Outliers Detected	Number of records with outlier values for specific numeric fields.
Negative Values Found	Number of records with invalid negative values (e.g., negative medal counts).
Quality Metric	Consolidated quality metrics for nulls, duplicates, and validation errors.
Pipeline Status	Status of the pipeline (e.g., Success, Failed).
Processed Date	Date when the data was processed.

Data Zones

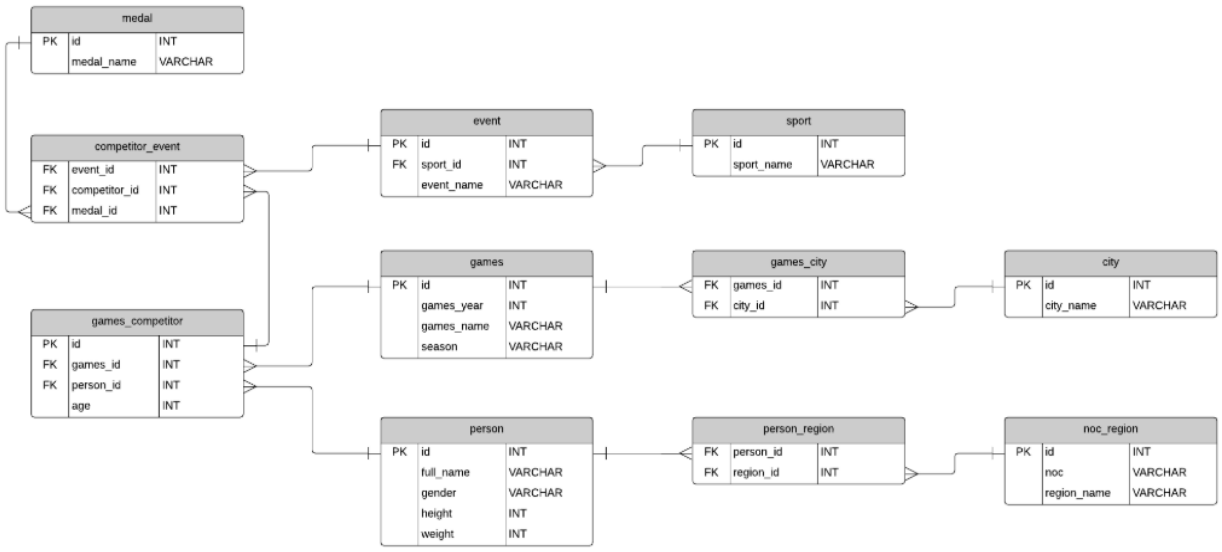
The dataset undergoes the following transformations through the pipeline:

1. **Bronze Zone:** Stores the raw data ingested from Azure Blob Storage.
2. **Silver Zone:** Contains cleaned and standardized data after transformations (e.g., type casting, null handling).
3. **Gold Zone:** Contains aggregated data optimized for analytics and visualization.

4. Architectural Diagram:



5. ER Diagram:



6.How it Works:

1. *Data Ingestion*:
   - Raw CSV files from GitHub are ingested into Azure Data Lake using ADF pipelines.
2. *Data Quality Checks*:
   - Databricks notebooks validate the data with 10 quality checks (e.g., null checks, duplicate detection, range validation).
3. *Metrics Storage*:
   - Quality metrics are stored in CSV format in the Azure Data Lake for reporting.
4. *Alerts*:
   - Azure Monitor triggers alerts for pipeline failures or quality anomalies.
5. *Reporting*:
   - Power BI connects to Azure Data Lake to visualize metrics and trends.

7. Execution Overview:

Step 1: Ingestion (Bronze Zone)

- The raw data file (Global_Superstore.csv) is uploaded to **Azure Blob Storage**.
- Azure Data Factory (ADF) pipelines copy the file into **Azure Data Lake Storage Gen2 (ADLS)**.
- The ingestion tasks involve mounting the data in **Databricks** for further processing.

Step 2: Transformation (Silver Zone)

- Using Databricks Notebooks, the raw data is cleaned and transformed into structured data:
  - Data types are cast appropriately (e.g., sales as float, quantity as int).
  - Columns like Profit Margin and High Discount are calculated.
  - Null values and invalid rows are removed.

Step 3: Analytical Zone (Gold Zone)

- The cleaned and aggregated data in the Silver Zone is used to:
  - Compute sales and profit by region, category, and year.
  - Identify high-profit orders and discount trends.
- The processed data is used in Delta format for visualizations and insights.

Pipeline Overview:

1. *Ingestion Pipeline*:
   - Extracts data from GitHub and loads it into the Raw Zone.
2. *Transformation Pipeline*:
   - Executes Databricks notebooks for quality checks.
   - Processes data into the Processed Zone.
3. *Metrics Pipeline*:
   - Consolidates and stores quality metrics in the Metrics Zone.
4. *Monitoring Pipeline*:
   - Tracks pipeline performance and triggers alerts for anomalies.

8. Resources Used for the project:

-*Azure Data Factory*: For orchestration of data pipelines.

- *Azure Data Lake Storage*: To store raw, processed, and metrics data.

- *Azure Databricks*: For data quality checks and advanced transformations.

- *Azure Monitor*: For pipeline and quality monitoring with alerts.

- *Power BI*: For reporting and visualization of quality metrics.

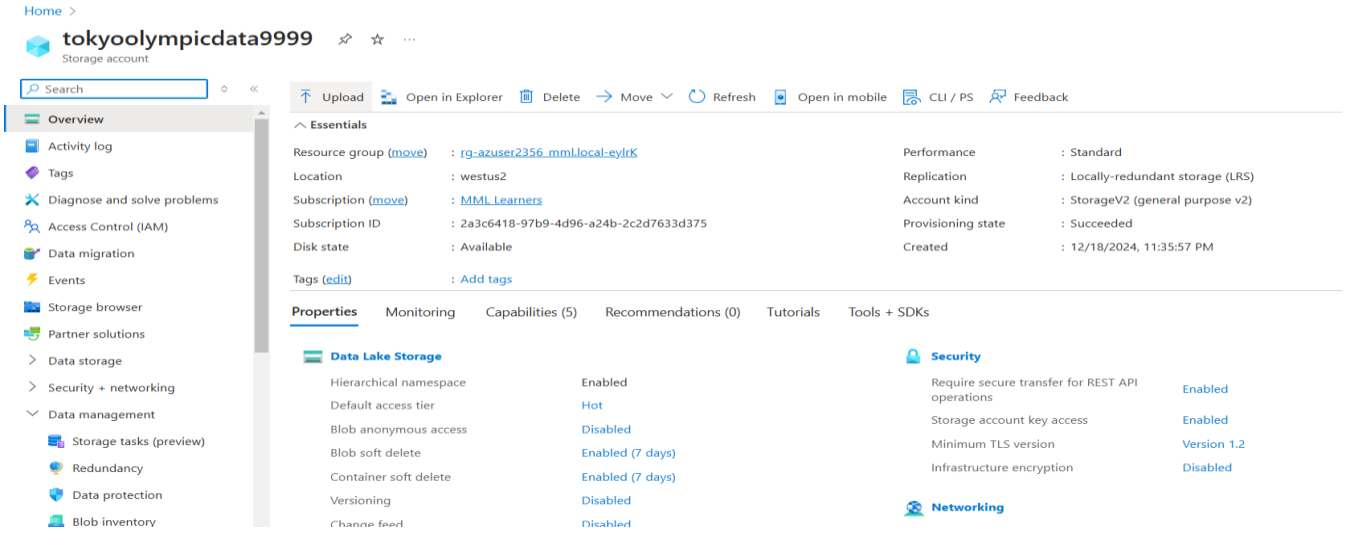
- *GitHub*: Source repository for raw CSV files.

9. Project Requirements:

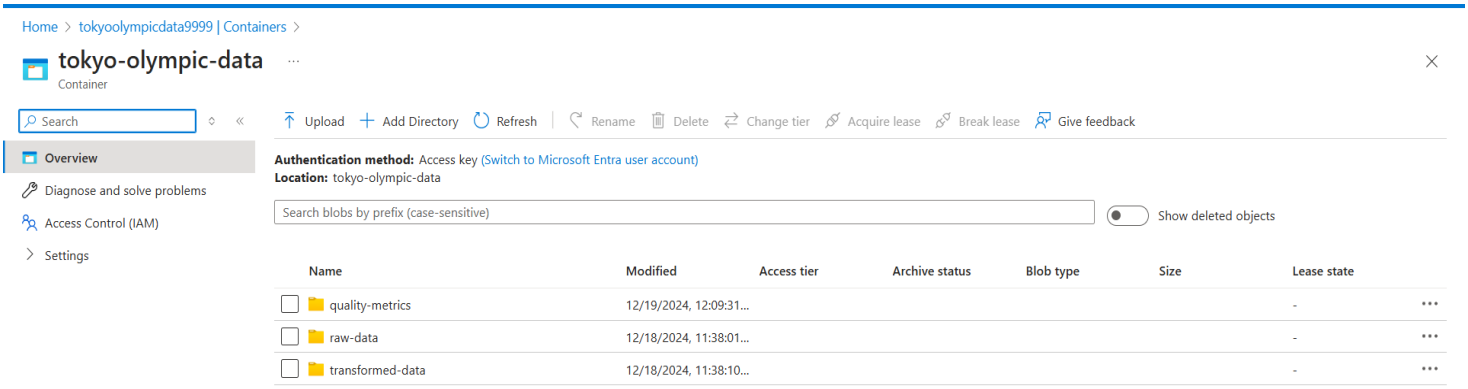
-*Azure Subscription* with access to ADF, Data Lake Storage, and Databricks.
- *Data Sources*:
  - Tokyo Olympics dataset in CSV format.
- *Tools and Software*:
  - Power BI Desktop for report creation.
- *Permissions*:
  - Access to Azure resources (storage, compute, and monitoring).

10. Tasks Performed:

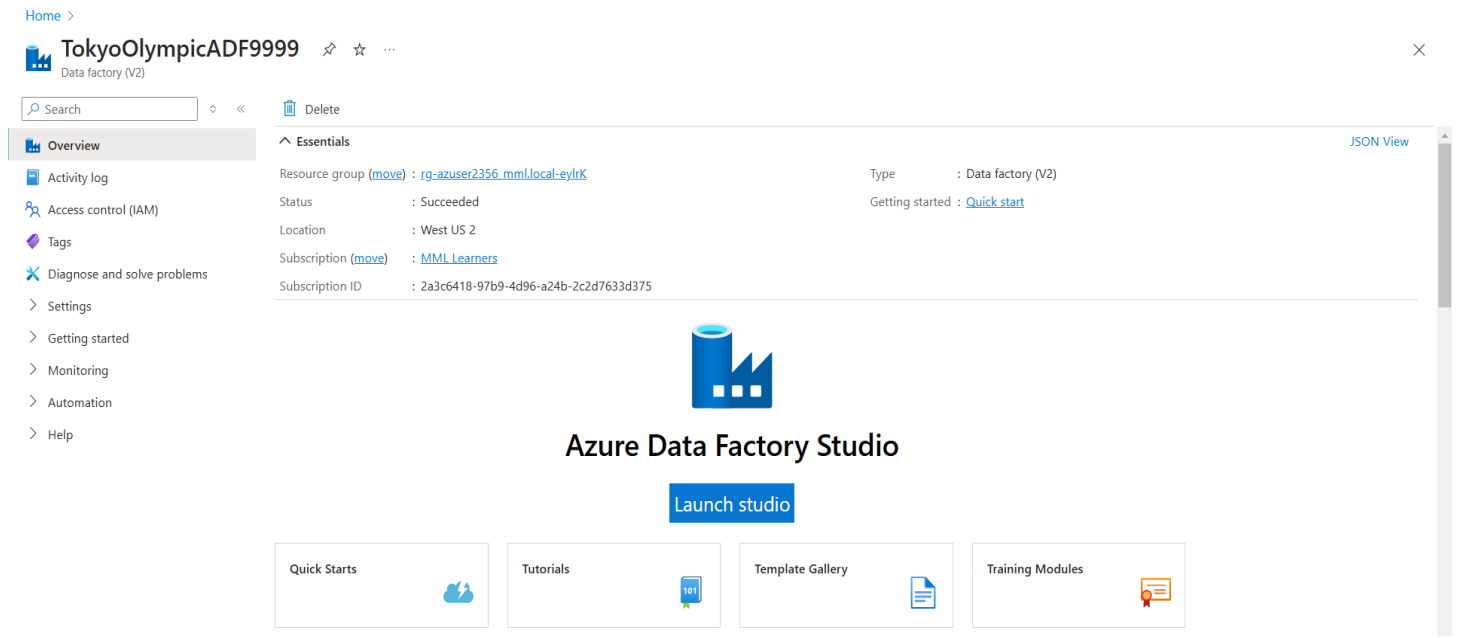
Creating an Data Lake storage Account:



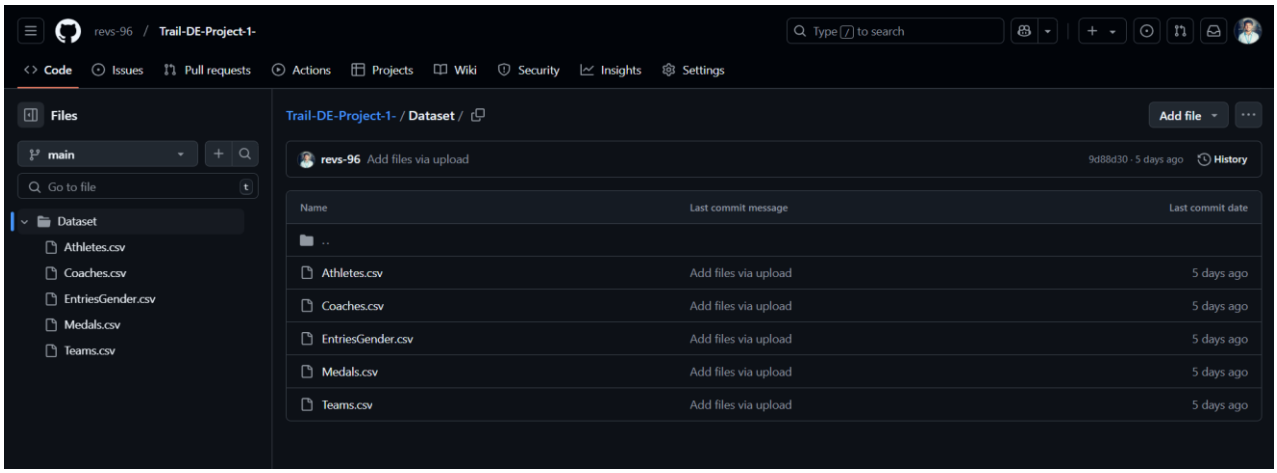
Creating container and directories in delta lake storage account:



Creating ADF Account:

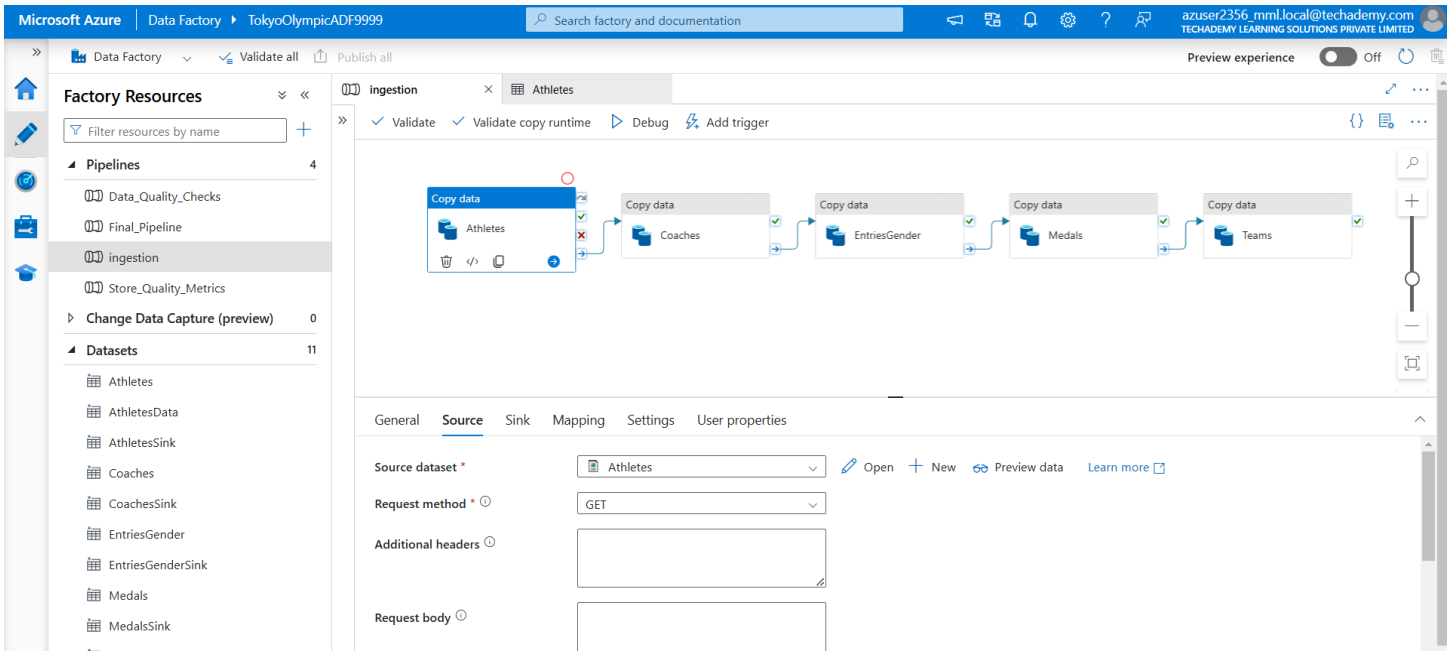


Uploading CSV Files into github to access through pipeline:

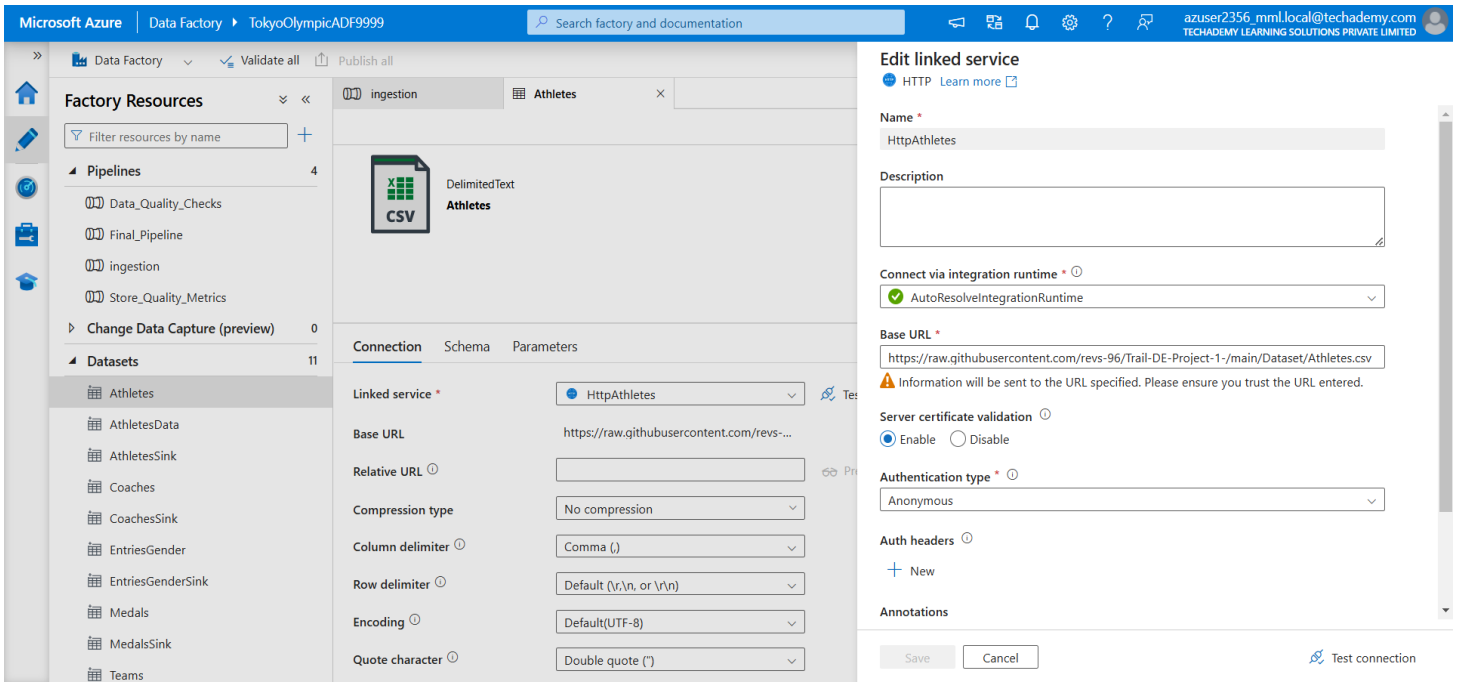


Ingesting data from github(Source) to Data lake storage(sink):

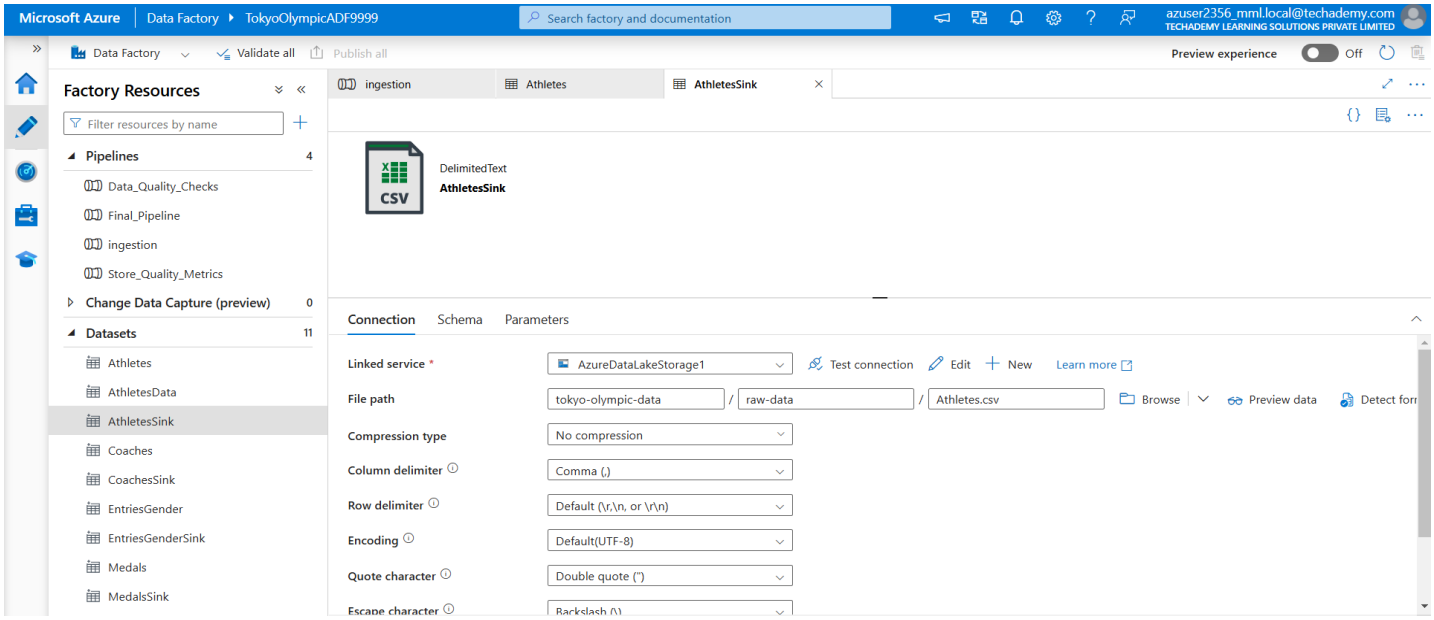
INGESTION PIPELINE



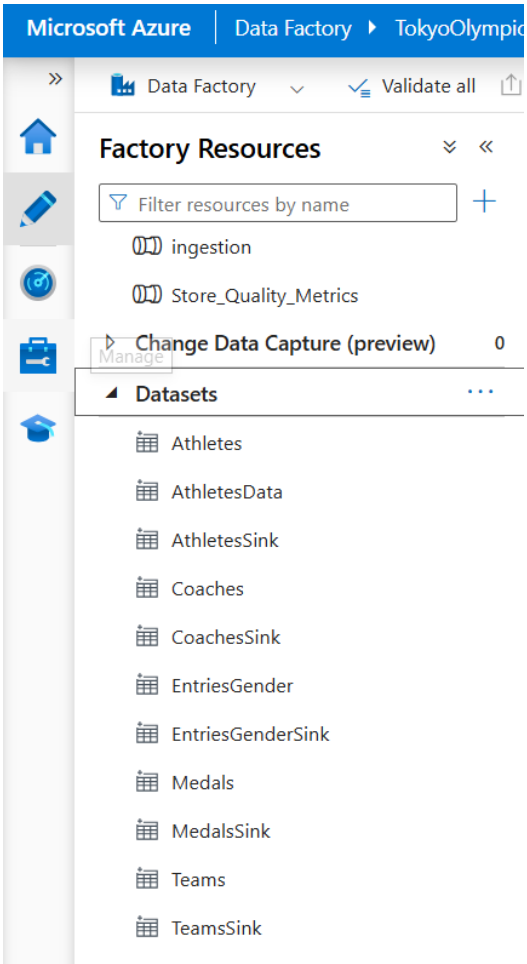
Source:



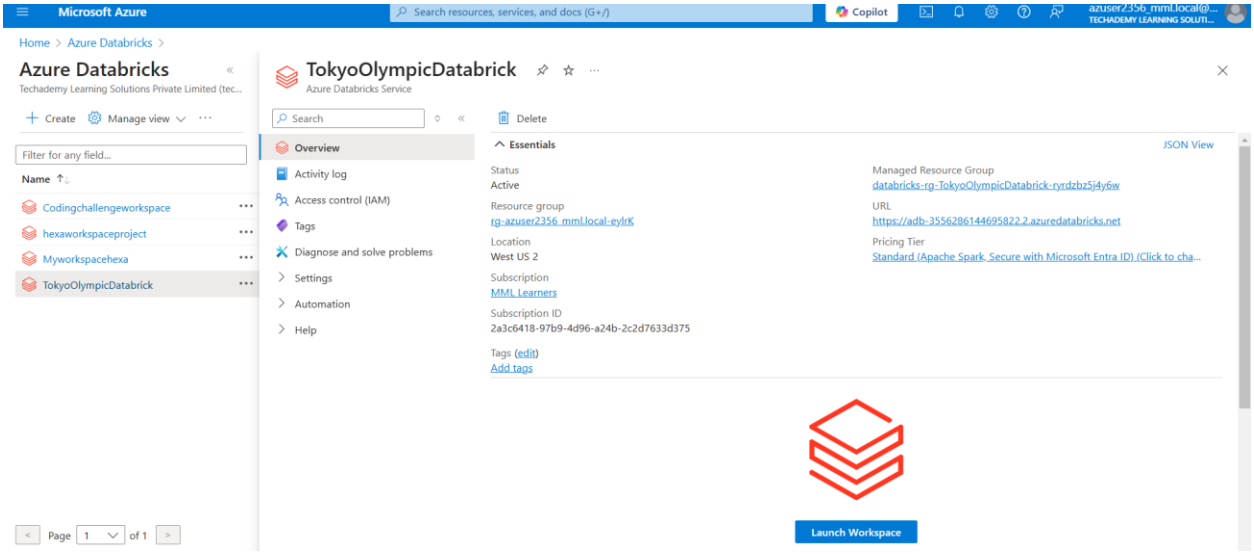
Sink:



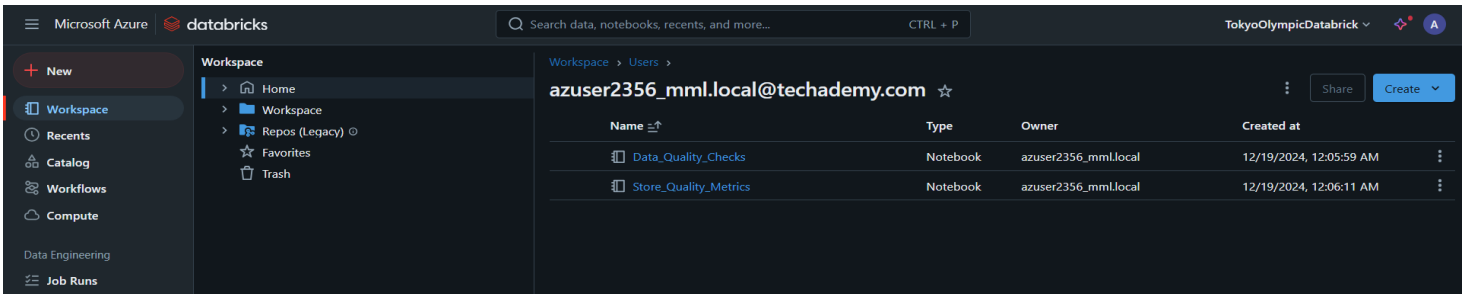
Total Datasets Imported:



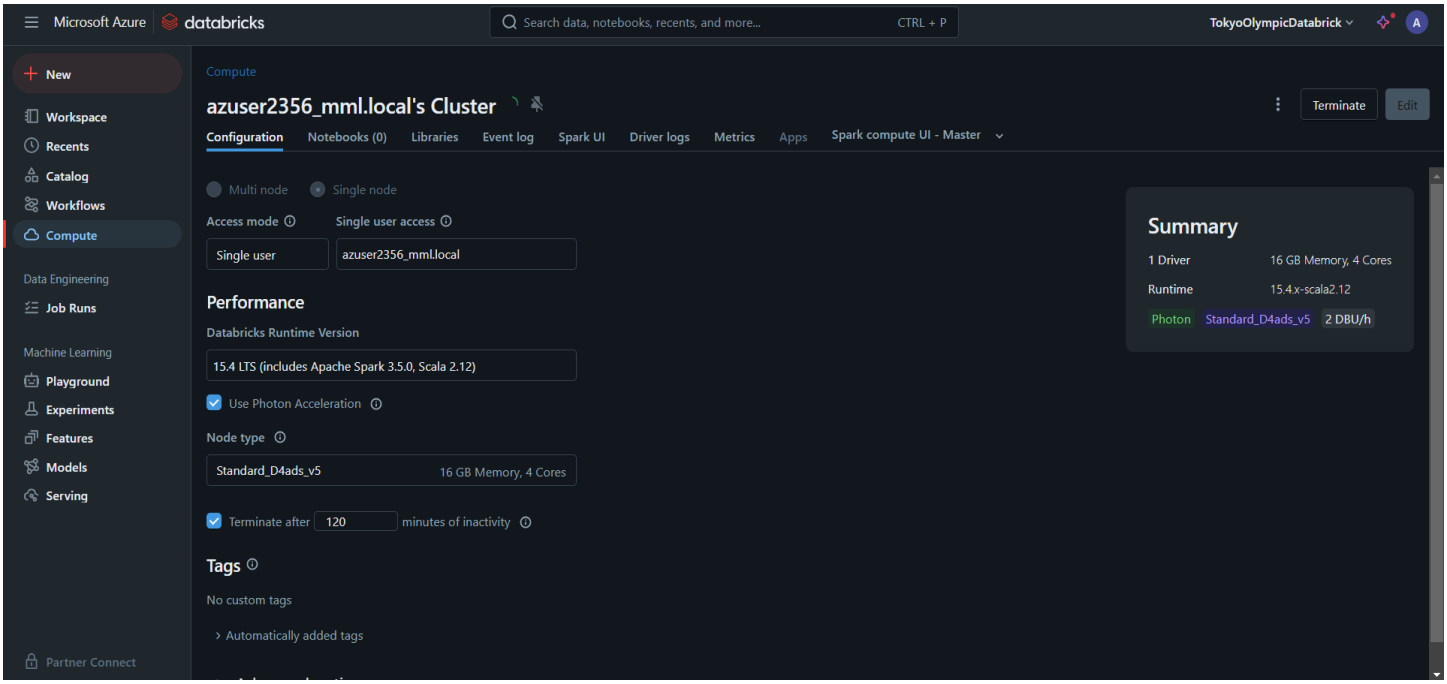
Tokyo Olympics Databricks workspace:



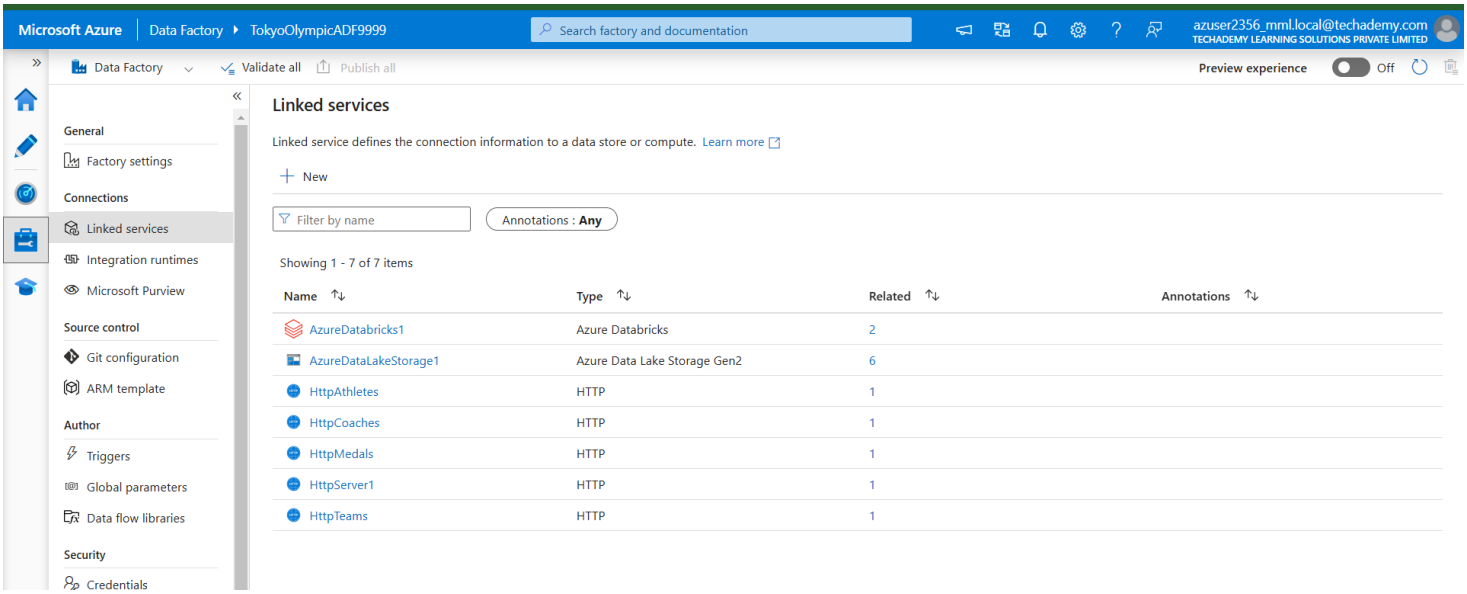
Databricks notebooks:



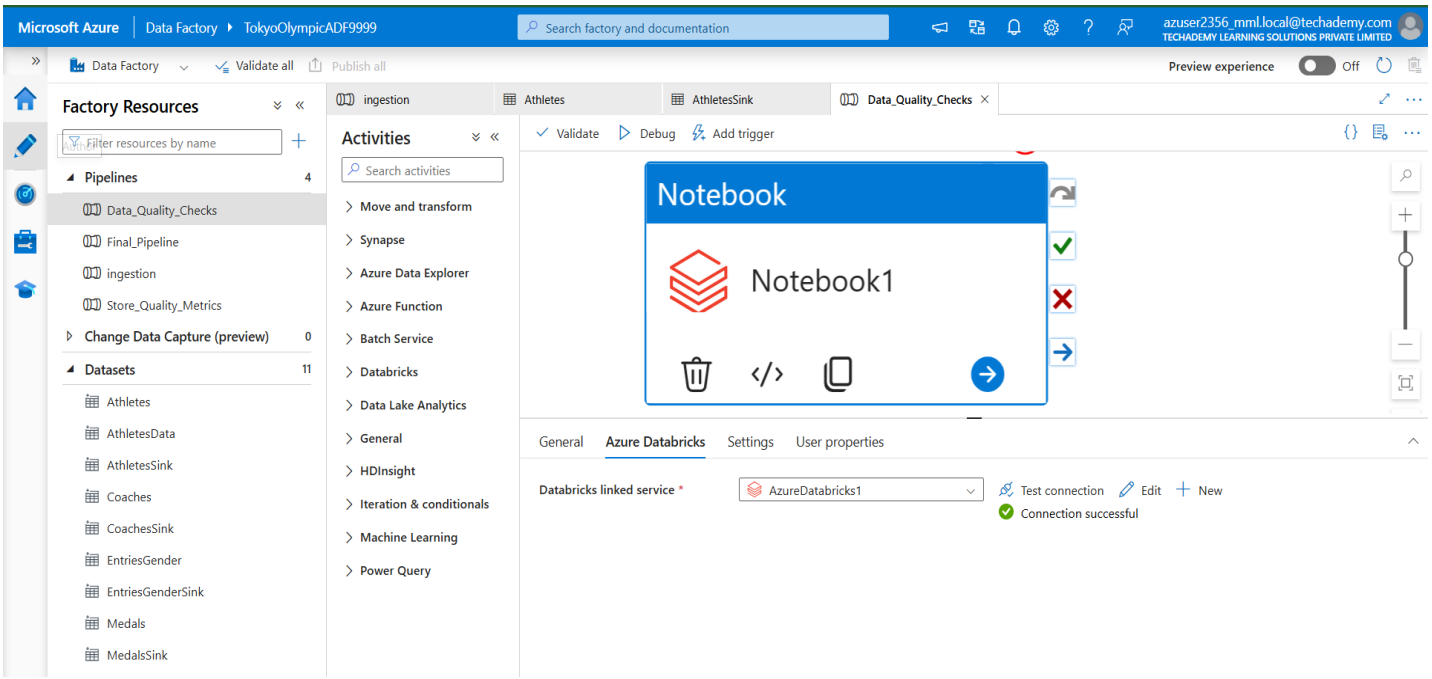
Cluster Creation:



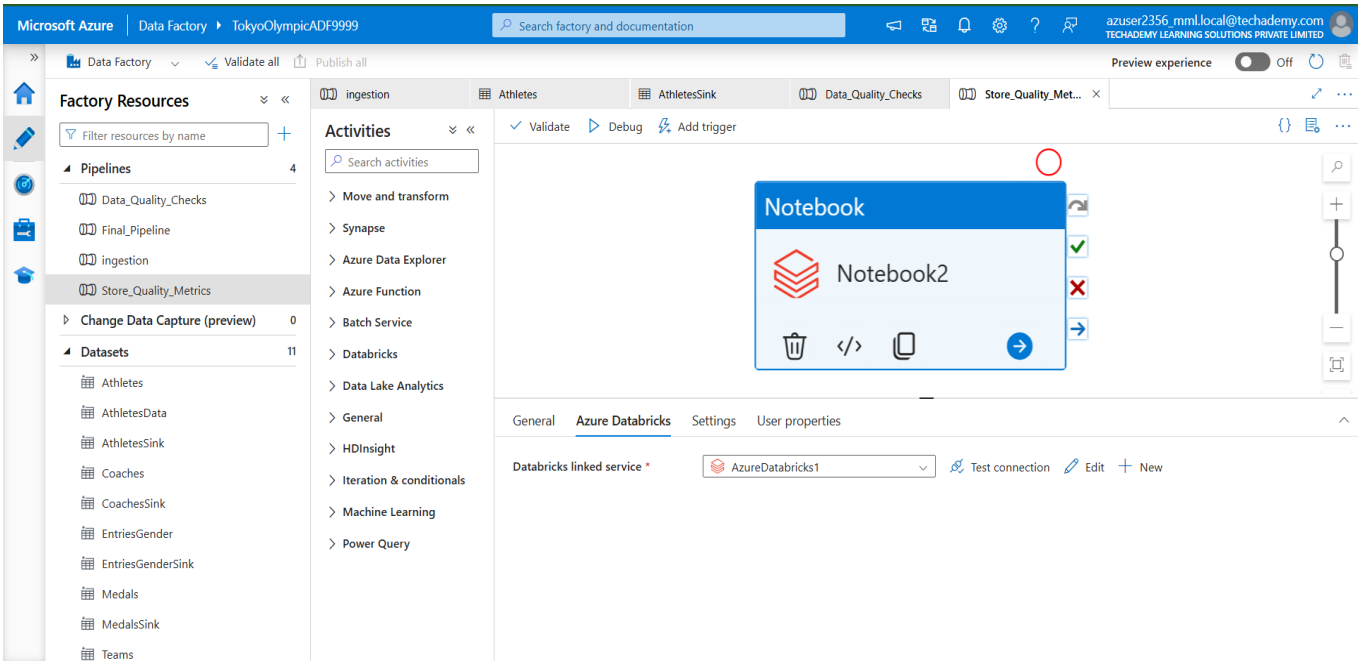
Linked Services in ADF:



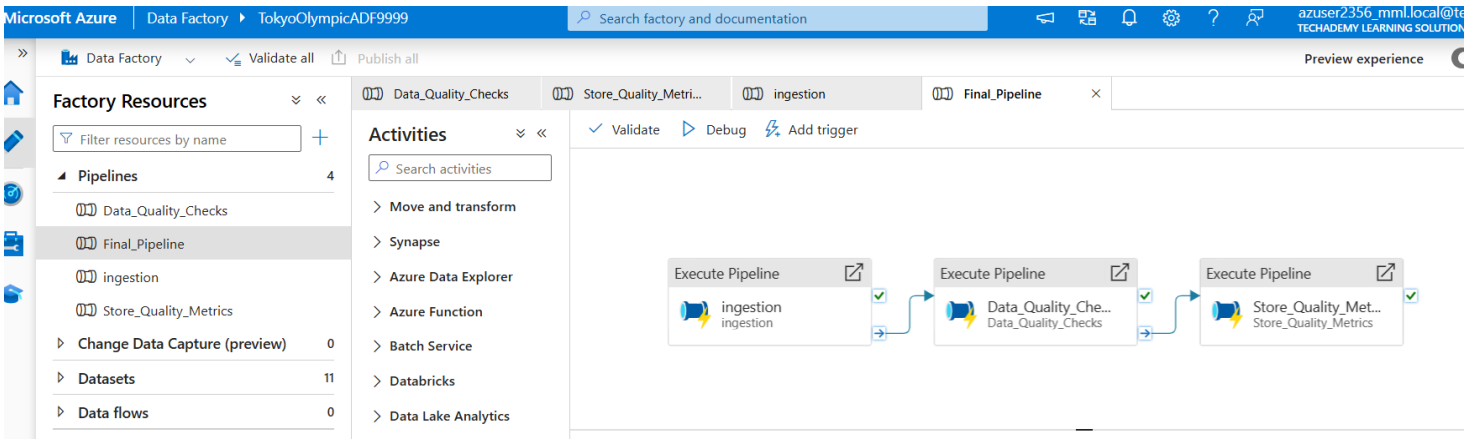
Data Quality check pipeline:



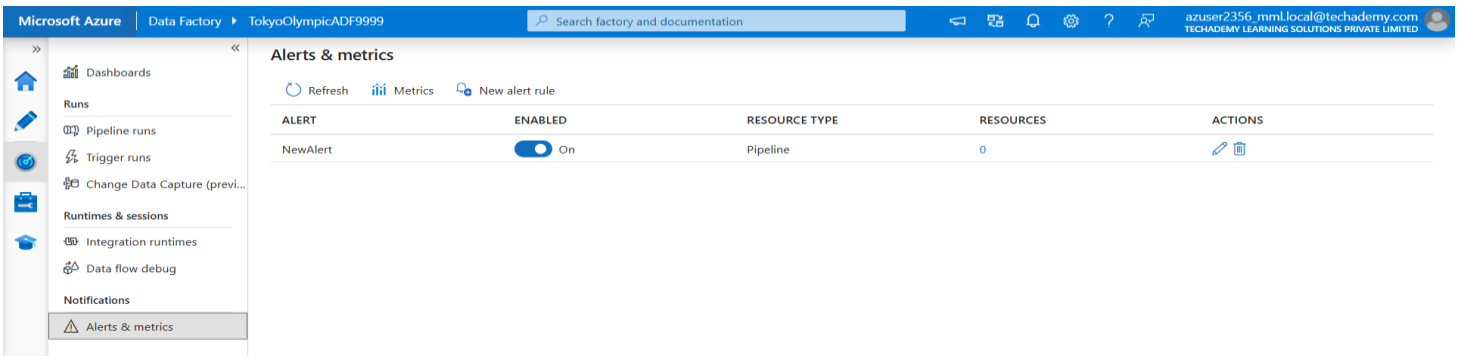
Store Quality metrics Pipeline:



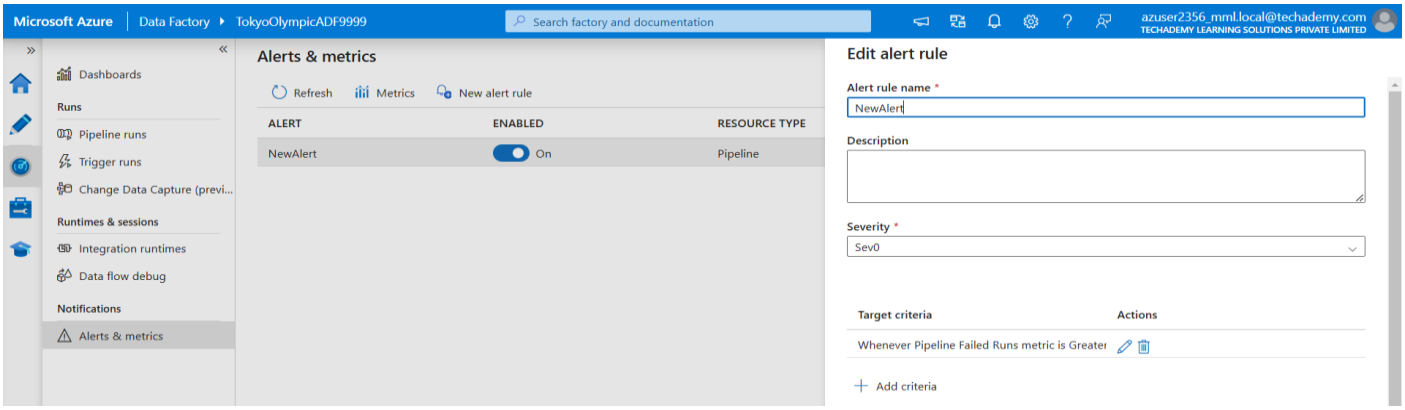
Final Execution Pipeline:



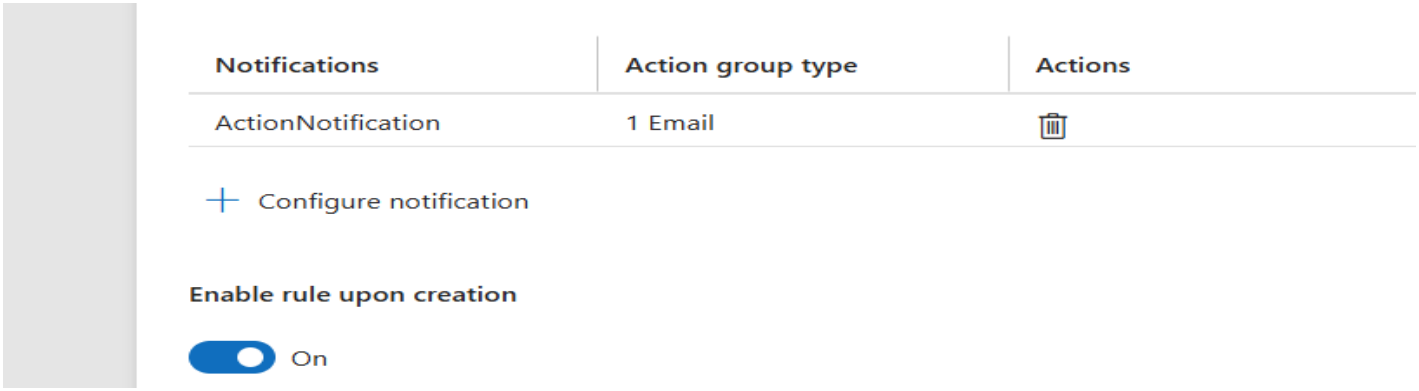
Alerts and Metrics:



Target Criteria:



Configured Email notification:



Submitted by:

Aathirainathan P